DATA PARSING AND TOKENIZING APPARATUS, METHOD AND PROGRAM

ABSTRACT OF THE DISCLOSURE

Apparatus for parsing and tokenizing a data stream comprises: a storage component to store a history buffer containing an unencoded version of a previously encoded string; a comparison component to compare a string from the input data stream with the unencoded version of at least one previously encoded string; a second storage component store: an indicator that at least two matches were found by the first comparison component, and tokens corresponding to the matches; a summing component to sum potential token lengths to provide total potential token lengths; a second comparison component to compare total potential token lengths; a selection component to select a match corresponding to a shortest total token length to represent the string from said input data stream; and an emitting component for emitting tokens representing the match corresponding to the shortest total token length. The tokens may be used in, for example, compression or encryption.